DETERMINATIVE DEGREE AND NUCLEOTIDE SEQUENCE ANALYSIS BY TRIANDERS

Diana Duplij [1)]
Institute of Molecular Biology and Genetics, 150 Zabolotny Str.
Kiev 03143, Ukraine

Steven Duplij [2)]
Kharkov National University, 4 Svoboda. Sq.
Kharkov 61077, Ukraine

February 9, 2008

Abstract

A new version of DNA walks, where nucleotides are regarded unequal in their contribution to a walk is introduced, which allows us to study thoroughly the ”fine structure” of nucleotide sequences. The approach is based on the assumption that nucleotides have an inner abstract characteristics, the determinative degree, which reflects phenomenological properties of genetic code and is adjusted to nucleotides physical properties. We consider each position in codon independently, which gives three separate walks being characterized by different angles and lengths, and such an object is called triander which reflects the “strength” of branch. A general method of identification of DNA sequence ”by triander”, which can be treated as a unique ”genogram”, ”gene passport” is proposed. The two- and three-dimensional trianders are considered. The difference of sequences fine structure in genes and the intergenic space is shown. A clear triplet signal in coding locuses is found which is absent in the intergenic space and is independent from the sequence length. The topological classification of trianders is presented which can allow us to provide a detail working out signatures of functionally different genomic regions.

[1] E-mail: duplijd@mail.ru

[2] E-mail: Steven.A.Duplij@univer.kharkov.ua; WWW: http://www.math.uni-mannheim.de/~duplij

1 Introduction

The genomic DNA sequence analysis using wide range of statistical methods [1, 2, 3, 4, 5, 6, 7] and various symmetry investigations [8, 9, 10, 11, 12, 13] is an extremely important tool in extracting hidden information about the dynamic process of evolution, especially after the availability of fully sequenced genomes [14]. One of the most promising approaches is the DNA walks method [15, 16, 17] (firstly introduced by Azbel [18]) or genomic landscapes [19], which is based on mapping of a sequence into one-, two- or multidimensional metric space according to various specific rules. In brief, while drawing a DNA walk, the corresponding mappings assign a direction/unit vector to each nucleotide, to dinucleotide or to purine (pyrimidine). The resulting broken lines endow a visual presentation to a formal sequence of 4 symbols, where inhomogeneous regions, fluctuations, "patches" etc. [20] are immediately seen. A modification of the DNA walks method deals with each position in codons independently, which gives three separate broken lines being characterized by different angles and lengths [21], where also addition and subtraction of DNA walks were considered [22].

Here we introduce a new version of DNA walks, where all 4 nucleotides are regarded unequal in the sense that they give contribution to a walk differing not only by direction, but also by module. It follows from the assumption [23] that nucleotides have an inner abstract characteristics — the determinative degree [24] which reflects phenomenological properties of genetic code and is adjusted to nucleotides physical properties.

2 Genetic code redundancy, doublet matrix inner structure and determinative degree

As is well-known, the genetic code is a highly organized system [25] and has several general properties: triplet character, uniqueness, nonoverlapping, commaless, redundancy (degeneracy), which means that most amino acids can be specified by more than one codon [26, 27].

From 64 possible codons one can extract 16 families each defined by first two nucleotides. Let we denote a triplet (5'–1–2–3–3') by **XYZ**. Then the codon sense can be fully determined by first two nucleotides **X** and **Y** independently of third **Z**. There are 8 unmixed families (all 4 codons encode the same amino acid), and 8 mixed families for which several patterns of assignment exist, in 6 of the latter the pyrimidine codons ($\mathbf{Z} = \mathbf{C}, \mathbf{U}$) determine one amino acid, and the purine codons ($\mathbf{Z} = \mathbf{A}, \mathbf{G}$) determine other ones or termination signals (in one family). It was found that two third part of all DNA bases are identical for all organisms for the sake of first two nucleotides in a triplet, and variability of DNA composition is given by the third base [28, 26].

All 16 doublets **XY** can be presented as the canonical matrix [29]

$$\begin{array}{ccccccc}
 & & & \mathbf{CC} & & & \\
 & & \mathbf{GC} & & \mathbf{CG} & & \\
 & \mathbf{CU} & & \mathbf{GG} & & \mathbf{CU} & \\
\mathbf{AC} & & \mathbf{UG} & & \mathbf{GU} & & \mathbf{CA} \\
 & \mathbf{AG} & & \mathbf{UU} & & \mathbf{GA} & \\
 & & \mathbf{AU} & & \mathbf{UA} & & \\
 & & & \mathbf{AA} & & &
\end{array} \tag{1}$$

called the “rhombic code” [30, 31]. They are grouped together in 2 octets distinguished by ability of amino acid determination: 8 doublets **CC**, **AC**, **GC**, **CU**, **GU**, **UC**, **CG**, **GG** determine amino acid independently of third base (upper part in (1)), and so they can be called “strong”, and other 8 doublets **AA**,**AU**, **UU**, **CA**, **GA**, **UG**, **AG, UA** (lower part in (1)) for which third base determines content of codons can be called “weak” ones [29, 32]. The “strong” set of doublets has the following relative content $\mathbf{C} : \mathbf{G} : \mathbf{U} : \mathbf{A} = 7 : 5 : 3 : 1$, while the “weak” set has the reverse content $\mathbf{C} : \mathbf{G} : \mathbf{U} : \mathbf{A} = 1 : 3 : 5 : 7$ [33]. Note that there is only one **A** in the “strong” octet, and one **C** in “weak” octet, and all 4 doublets with $\mathbf{Y} = \mathbf{C}$ completely determine amino acid, but only 2 doublets with $\mathbf{Y} = \mathbf{G}$ and $\mathbf{Y} = \mathbf{U}$ completely determine it, while doublets with $\mathbf{Y} = \mathbf{A}$ never determine amino acid. Thus, 4 nucleotides can be arranged in descending order **C, G, U, A** by their determinative ability (“strength”) [29, 34].

We introduce a numerical characteristics of the empirical “strength” — *determinative degree* of nucleotide $\mathbf{d_X}$ in the following way

$$\begin{array}{cccc}
\text{Pyrimidine} & \text{Purine} & \text{Pyrimidine} & \text{Purine} \\
\mathbf{C} & \mathbf{G} & \mathbf{T/U} & \mathbf{A} \\
\mathbf{d_C = 4} & \mathbf{d_G = 3} & \mathbf{d_{T/U} = 2} & \mathbf{d_A = 1} \\
\text{very “strong”} & \text{“strong”} & \text{“weak”} & \text{very “weak”} \\
\textit{completely} & \textit{in 2 cases} & \textit{in 2 cases} & \textit{never}
\end{array} \tag{2}$$

which allows us to make transition from qualitative to quantitative description of genetic code structure [23, 35].

We use the notation **T/U**, because genetic code is read from mRNA, and so we will not differentiate their determinative ability (“strength”) in what follows.

Let us present four bases (2) as the vector-column

$$\mathbb{V} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \mathbf{V}_3 \\ \mathbf{V}_4 \end{pmatrix} = \begin{pmatrix} \mathbf{C}^{(4)} \\ \mathbf{G}^{(3)} \\ \mathbf{T}^{(2)} \\ \mathbf{A}^{(1)} \end{pmatrix} \tag{3}$$

and the corresponding the vector-row

$$\mathbb{V}^T = \left(\begin{array}{cccc} \mathbf{C}^{(4)} & \mathbf{G}^{(3)} & \mathbf{T}^{(2)} & \mathbf{A}^{(1)} \end{array} \right). \tag{4}$$

where the upper index for nucleotide denotes determinative degree. We make the exterior product of vector-column (3) and vector-row (4) as follows [23, 24]

$$\mathbb{M} = \mathbb{V} \times \mathbb{V}^T = \left(\begin{array}{cccc} \mathbf{C}^{(4)}\mathbf{C}^{(4)} & \mathbf{C}^{(4)}\mathbf{G}^{(3)} & \mathbf{C}^{(4)}\mathbf{T}^{(2)} & \mathbf{C}^{(4)}\mathbf{A}^{(1)} \\ \mathbf{G}^{(3)}\mathbf{C}^{(4)} & \mathbf{G}^{(3)}\mathbf{G}^{(3)} & \mathbf{G}^{(3)}\mathbf{T}^{(2)} & \mathbf{G}^{(3)}\mathbf{A}^{(1)} \\ \mathbf{T}^{(2)}\mathbf{C}^{(4)} & \mathbf{T}^{(2)}\mathbf{G}^{(3)} & \mathbf{T}^{(2)}\mathbf{T}^{(2)} & \mathbf{T}^{(2)}\mathbf{A}^{(1)} \\ \mathbf{A}^{(1)}\mathbf{C}^{(4)} & \mathbf{A}^{(1)}\mathbf{G}^{(3)} & \mathbf{A}^{(1)}\mathbf{T}^{(2)} & \mathbf{A}^{(1)}\mathbf{A}^{(1)} \end{array} \right). \tag{5}$$

It is remarkable that the matrix $\mathbb{M}$ (5) fully *coincides* with the canonical matrix of doublets (1), if and only if the vector $\mathbb{V}$ has the determinative degree order **C, G, U, A** (2). Although there are 4!=24 possibilities to place 4 bases in row, but all others except one presented in (2) do not reflect phenomenological properties of genetic code. It follows that the intuitive "rhombic code" and genetic vocabulary [29, 30, 31] have their own inner abstract structure uniquely defined by exterior product of special vectors (3). This ordering is also adjusted to the schemes [36, 37], also (partially) with half time of nucleotide substitution under mutational pressure [38] and the nucleotides information weights [39]. Indeed these facts allows us to introduce the determinative degree, as an *abstract variable* being a numerical measure of nucleotide difference in ability to determine sense of codon [23, 24].

Analogous model for the triplet genetic code can be constructed using triple exterior product in the same way [23]. We dispose the doublet matrix $\mathbb{M}$ on the **XY** plane and multiply it on the vector-column $\mathbb{V}$ (3) disposed along **Z** axis, i.e. we construct the triple exterior product

$$\mathbb{K} = \mathbb{V} \times \mathbb{M}. \tag{6}$$

Thus we obtain three-dimensional matrix over set of all triplets, and, since each codon (except three terminal ones) corresponds to an amino acid, that can be treated as a *cubic matrix model of the genetic code* [23].

3 Determinative degree and nucleotide properties

The connection bulk DNA structure and various properties of nucleotides was studies in [40, 41]. It is well-known that by chemical structure the 4 nitrous bases can be divided into:

1) purine (**A**,**G**) and pyrimidine (**C**,**T**);

2) having amino (**A**,**C**) group and (**G**,**T**) keto group;

3) making 3 (strong) hydrogen bonds (**C**,**G**) and 2 (weak) hydrogen bonds (**A**,**T**).

They give rise to 3 symmetry transformations:

1) Purine-pyrimidine symmetry

$$\begin{pmatrix} \mathbf{T}^{(2)} \\ \mathbf{A}^{(1)} \\ \mathbf{C}^{(4)} \\ \mathbf{G}^{(3)} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{C}^{(4)} \\ \mathbf{G}^{(3)} \\ \mathbf{T}^{(2)} \\ \mathbf{A}^{(1)} \end{pmatrix} = \mathcal{R}_{pur}\mathbb{V}; \tag{7}$$

2) Amino-keto symmetry

$$\begin{pmatrix} \mathbf{A}^{(1)} \\ \mathbf{T}^{(2)} \\ \mathbf{G}^{(3)} \\ \mathbf{C}^{(4)} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{C}^{(4)} \\ \mathbf{G}^{(3)} \\ \mathbf{T}^{(2)} \\ \mathbf{A}^{(1)} \end{pmatrix} = \mathcal{R}_{amino}\mathbb{V}; \tag{8}$$

3) Complementary symmetry (leaving the double helix invariant)

$$\begin{pmatrix} \mathbf{G}^{(3)} \\ \mathbf{C}^{(4)} \\ \mathbf{A}^{(1)} \\ \mathbf{T}^{(2)} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{C}^{(4)} \\ \mathbf{G}^{(3)} \\ \mathbf{T}^{(2)} \\ \mathbf{A}^{(1)} \end{pmatrix} = \mathcal{R}_{compl}\mathbb{V}, \tag{9}$$

where the even (because determinant is +1) permutation matrices $\mathcal{R}_{pur}, \mathcal{R}_{amino}, \mathcal{R}_{compl}$ satisfy

$$\mathcal{R}_{pur}\mathcal{R}_{amino}\mathcal{R}_{compl} = \mathcal{I},$$

and two of them, e.g. $\mathcal{R}_{pur}$, $\mathcal{R}_{compl}$ together with the identity matrix $\mathcal{I}$ form the dihedral group D_2 which is the symmetry group of the dihedron, or regular double-pyramid, with vertices on the unit-sphere (see e.g. [42]). Another representation of this group by 3×3 rotational matrices is called a DNA group [43].

The difference in the number of hydrogen bonds causes the different interaction with its complementary nucleotide: each “strong” nucleotide (**C** and **G**) has 3 bonds and the energy of **C**-**G** interaction is -2.4 kkal/mol, and each “weak” nucleotide (**T** and **A**) has only 2 bonds and the energy of **A**-**T** interaction is -1.2 kkal/mol [26]. Therefore each base has its own properties and so dividing them into only 2 groups is not sufficient.

We then can search whether the ordering (2) is adjusted to some physical properties of nucleotides.

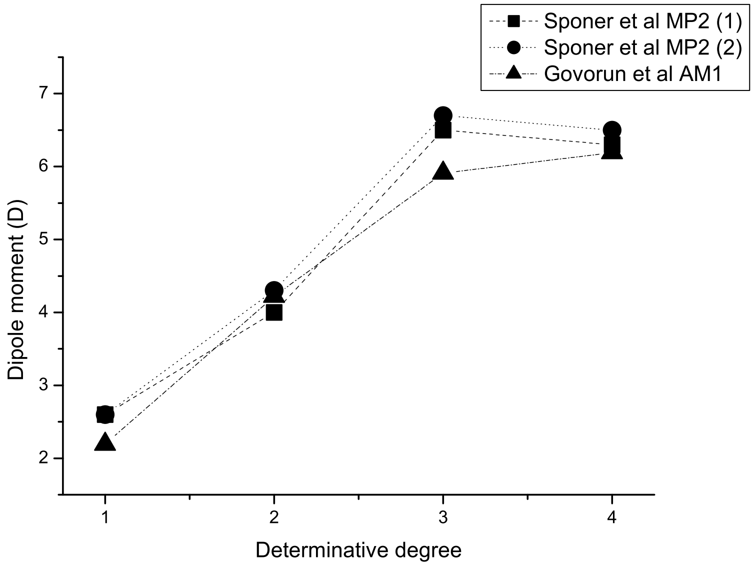

Figure 1: Dipole moment of DNA bases calculated by methods AM1 [41] (triangles) and two modifications of MP2 [44] (squares and circles). The corresponding linear fits are: $D_{\mathrm{AM1}} = 1.21 + 1.37\mathbf{d}_x$ ($R = 0.96$); $D_{\mathrm{MP2(1)}} = 1.45 + 1.36\mathbf{d}_x$ ($R = 0.93$); $D_{\mathrm{MP2(2)}} = 1.5 + 1.41\mathbf{d}_x$ ($R = 0.93$).

First we observe that the dipole moment of bases is proportional to the determinative degree as it is shown on Fig. 1.

Then we see that the weight of hydration sites for bases is also proportional to the determinative degree Fig. 2.

We can conclude that the determinative degree reflects not only redundancy of genetic code in the third position, but also connected with some energetic properties of bases themselves.

4 Trianders and their characteristics

It can be assumed that the phenomenological properties of genetic code and inequality of bases (reflected in (2)) will become apparent in real nucleotide sequences. Here we use the introduced determinative degree to build a new kind of sequence analysis based on some special modification of DNA walks method [18, 19, 21, 17].

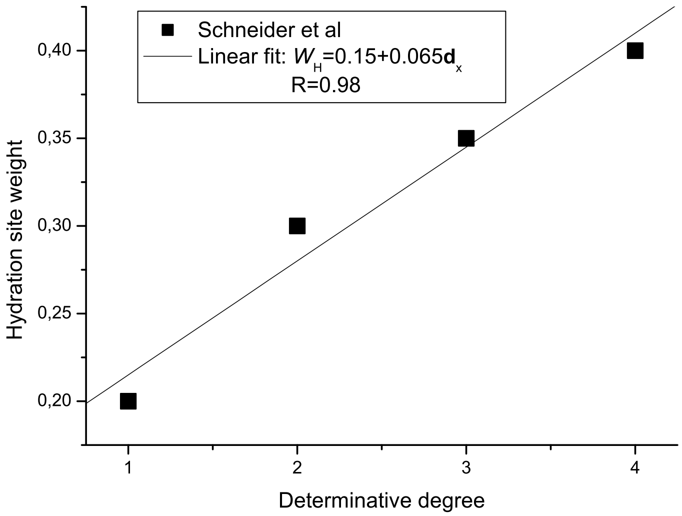

Figure 2: Weights of hydration site [45].

4.1 Triander construction

We embed a nucleotide sequence into the two-dimensional *determinative degree space* (DD plane) in the following way. The axis assignment corresponds to the value of nucleotide determinative degree as

$$\text{Axis } x\text{: } \{\mathbf{A}\} = (-1,0)\,;\{\mathbf{T}\} = (+2,0)\,,$$
$$\text{Axis } y\text{: } \{\mathbf{G}\} = (0,-3)\,;\{\mathbf{C}\} = (0,+4)\,.$$

Moving along a sequence produces a walk in the determinative degree space which we call a *determinative degree walk*. In general, a current point on DD plane after i steps is determined by the coordinates

$$x_i^{DD} = \mathbf{d_T} n_{\mathbf{T}}(i) - \mathbf{d_A} n_{\mathbf{A}}(i)\,, \tag{10}$$
$$y_i^{DD} = \mathbf{d_C} n_{\mathbf{C}}(i) - \mathbf{d_G} n_{\mathbf{G}}(i)\,, \tag{11}$$

where $n_{\mathbf{X}}(i)$ is cumulative quantity of nucleotide $\mathbf{X}$ after i steps and $\mathbf{d_X}$ is the determinative degree of nucleotide $\mathbf{X}$. The standard DNA walks [17] (genome landscapes [19]) have all $\mathbf{d_X} = 1$ in (10)–(11), i.e.

$$x_i^{standard} = n_{\mathbf{T}}(i) - n_{\mathbf{A}}(i)\,, \tag{12}$$
$$y_i^{standard} = n_{\mathbf{C}}(i) - n_{\mathbf{G}}(i)\,. \tag{13}$$

The one-dimensional (purine/pyrimidine) DNA walks are defined by only one coordinate, while x is chosen as position, i.e.

$$x_i^{pp} = i, \tag{14}$$

$$y_i^{pp} = n_{\mathbf{C}}(i) + n_{\mathbf{T}}(i) - n_{\mathbf{A}}(i) - n_{\mathbf{G}}(i). \tag{15}$$

Therefore, while “purine/pyrimidine” DNA walks manifestly show the purine/pyrimidine imbalance, the standard DNA walks (12)–(13) applied for one strand show DNA asymmetry [46, 47] (violation of the Parity Rule 2 [48]), the determinative degree walk (10)–(11) visually shows “strength” imbalance in one strand.

Then we build 3 independent determinative degree walks beginning from 1st nucleotide with step 3 (due to the triplet structure of genetic code). In this way we obtain 3 broken lines (*branches*) starting from the point of origin, and each of them presents the determinative degree walk through the following nucleotide numbers:

1st branch goes through **1,4,7,10,13**... positions;

2nd branch goes through **2,5,8,11,14**... positions;

3rd branch goes through **3,6,9,12,15**... positions.

These 3 branches on the determinative degree plane are called *triander*.

If 1st letter corresponds to the first start codon nucleotide, then the triander branches represent nucleotide sets in three codon positions independently.

As distinct from previous versions of DNA walks in which all 4 nucleotides are regarded equivalent in the sense they give equal by module shifts, in our approach each nucleotide gives contribution different by module (which is taken equal to its determinative degree). So, despite we obtain at first sight isomorphic to [21] plot, trianders show not only quantitative composition and pure statistical laws of symbol strings, but also *reflect connection* between nucleotide sequences and inner phenomenological properties of genetic code and physicochemical properties of bases.

As an example of triander we will take the dystrophin gene which is the largest gene found in nature, measuring 2.4 Mb, and is responsible for Duchenne (DMD) and Becker (BMD) muscular dystrophies [49]. The dystrophin RNA is differentially spliced, producing a range of different transcripts, encoding a large set of protein isoforms. Dystrophin is a large, rod-like cytoskeletal protein which is found at the inner surface of muscle fibers. The triander for the dystrophin gene is presented on Fig. 3. For comparison we also show the triander for a shuffled sequence of the same nucleotide composition. Obviously the ideal triander for uniformly random sequence consists of 3 flowing together lines from the origin having 45 degrees slope. This line also corresponds to the symmetric sequence satisfying the Parity Rule 2 [48]: $N_{\mathbf{C}} = N_{\mathbf{G}}$, $N_{\mathbf{T}} = N_{\mathbf{A}}$. Such lines are presented on all triander plots below for normalization.

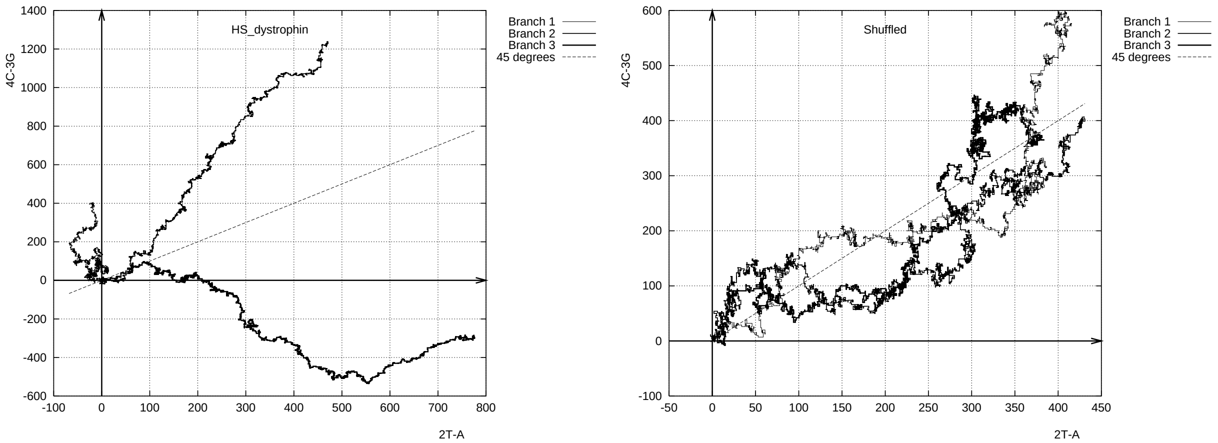

Figure 3: Triander for the Homo Sapiens dystrophin gene (left) and a shuffled sequence of the same nucleotide composition (right).

4.2 Determinative degree angle

An important visual characteristic of a triander is the slop of its branches, we call it determinative degree (DD) angle, which for a current point can be calculated by

$$\tan \alpha\,(i) = \frac{4n_{\mathbf{C}}\,(i) - 3n_{\mathbf{G}}\,(i)}{2n_{\mathbf{T}}\,(i) - n_{\mathbf{A}}\,(i)}.$$

Here and below $n_{\mathbf{X}}$ denotes cumulative quantity of nucleotide $\mathbf{X}$ for a given branch. Evidently, for uniformly random sequence or a symmetric sequence satisfying the Parity Rule 2 [48] the angle will be 45 degree (horizontal dashed line of the below plots), and so the difference from this value will say about nontrivial ordering. The plots of current values of α for the dystrophin gene and for a shuffled sequence of the same nucleotide composition are presented on Fig. 4.

We stress that trianders show not only quantitative composition, but allow us to find local motives in a more clear way, because different modules for nucleotides lead to less number of superposition and selfintersections. Also trianders more accurately reflect the tendency of the sequence as a whole similarly to DNA walks. Thus triander can be treated as a "picture", "genome passport" or "genogramma" of a given sequence.

If we remember that third base in codon has maximal redundancy, then 3rd branch of a triander gets a definite "physical sense". Let us assume that the determinative degree is an additive variable (which can be made in first approximation at least [23]), then 3rd branch can show current "strength" of sequence, that is the "bulk" ability to determine sense of codon. In this scheme other two branches can be treated as 3rd branch with shifted ORF (Open Reading Frame).

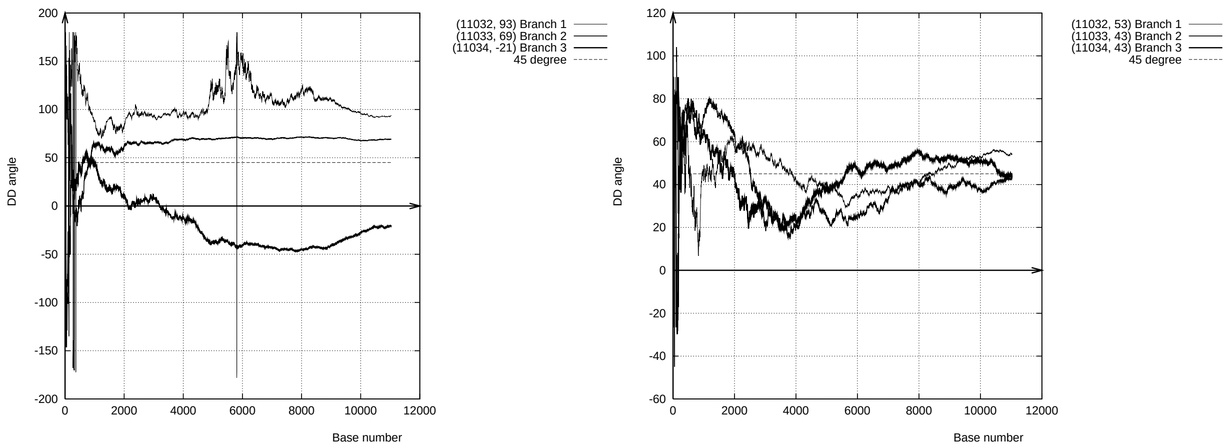

Figure 4: Current DD angle for triander of the Homo Sapiens dystrophin gene (left) and a shuffled sequence of the same nucleotide composition (right).

4.3 Euclidean and Manhattan distances

As the measure of the "sequence strength" we can choose length of the radius-vector from the origin to the current point of triander, i.e. the Euclidean distance

$$D_E\left(i\right)=\sqrt{\left(4n_{\mathbf{C}}\left(i\right)-3n_{\mathbf{G}}\left(i\right)\right)^2+\left(2n_{\mathbf{T}}\left(i\right)-n_{\mathbf{A}}\left(i\right)\right)^2}.$$

We can also use the Manhattan distance[3]

$$D_M\left(i\right)=\left|4n_{\mathbf{C}}\left(i\right)-3n_{\mathbf{G}}\left(i\right)\right|+\left|2n_{\mathbf{T}}\left(i\right)-n_{\mathbf{A}}\left(i\right)\right|,$$

which is the distance between two points measured along axes at right angles (see e.g. [50]).

In case of symmetric sequence (equal number of all nucleotides) at the step i the Euclidean and Manhattan distances are $D_E\left(i\right)=i/\sqrt{2}$ and $D_M\left(i\right)=i/2$ (which is shown by dashed lines on Figs. 5,6).

4.4 Visualization of the genetic code triplet nature

Now we make sure that triplet character of the genetic code can be seen directly from sequences representation by trianders. As an example we take gene of Homo sapiens Che-1 mRNA. We consider additionally analogs of trianders with different phases = 4,5,7. The result is presented on Fig. 7, from which it is seen that only the case phase = 3 provides nontrivial ordering leading to definite branches, that is we have clear visual presentation of the strong triplet signal.

[3] Also known as *rectilinear distance,* and it can be treated as the distance that would be traveled to get from one data point to the other if a grid-like path is followed (a car driving in a city laid out in square blocks, like Manhattan).

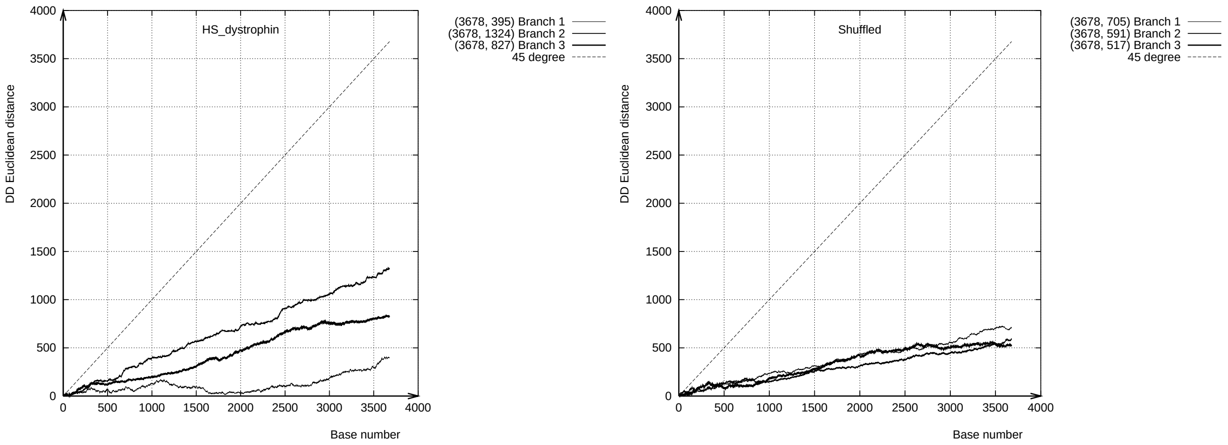

Figure 5: Current DD Euclidean distance for triander of the Homo Sapiens dystrophin gene (left) and a shuffled sequence of the same nucleotide composition (right).

In such a way one could search for higher phase statistical correlations and possible structures, if any, in nucleotide sequences.

4.5 Transformations of trianders

Here we illustrate how the symmetry transformations influence on the triander. As an example we take the Homo sapiens dystrophin from Fig. 3, and result of various symmetry transformations (7)–(9) and reversing the sequence is shown on Fig. 8.

We observe that the reverse triander is very similar to the original one on Fig. 3.

4.6 Three-dimensional trianders

The previously constructed two-dimensional trianders have the disadvantage form, because it is not clear, where in the sequence a given point is. To improve this we introduce three-dimensional trianders which are defined by the formula

$$x_i^{3D} = \mathbf{d_T} n_\mathbf{T}\left(i\right) - \mathbf{d_A} n_\mathbf{A}\left(i\right), \tag{16}$$

$$y_i^{3D} = \mathbf{d_C} n_\mathbf{C}\left(i\right) - \mathbf{d_G} n_\mathbf{G}\left(i\right), \tag{17}$$

$$z_i^{3D} = i, \tag{18}$$

which can be treated as mixing of one-dimensional and two-dimensional cases with taking into account the determinative degree. Then, any on the DD space structure

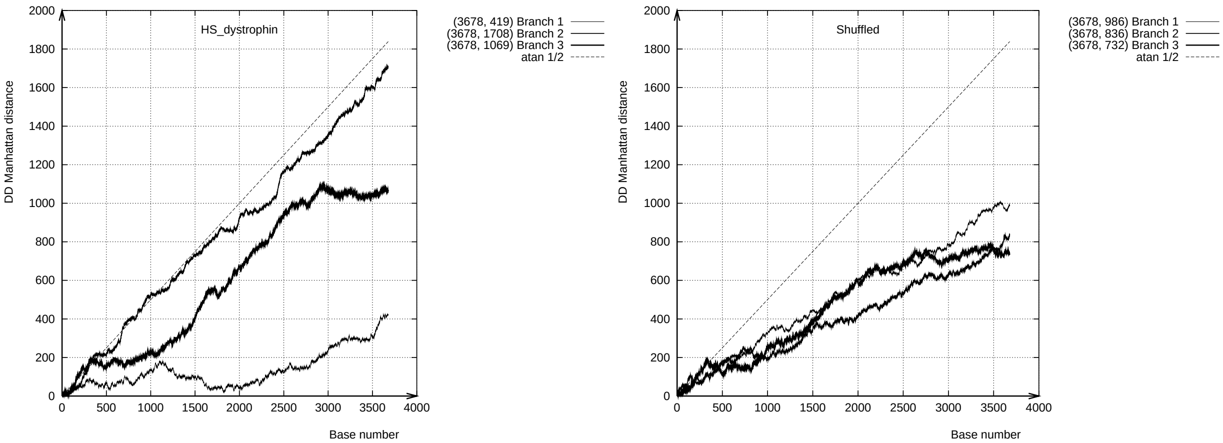

Figure 6: Current DD Manhattan distance for triander of the Homo Sapiens dystrophin gene (left) and a shuffled sequence of the same nucleotide composition (right).

can be definitely visually localized using vertical axis. On the Fig. 9 we show the three-dimensional triander of the Homo sapiens zinc finger and its shuffled version. All the graphs start from one point, the origin, and have different length (which can be simply calculated from (16)–(18)), characterizing them as a whole.

5 Topological classification of trianders

Here we propose the topological classification of trianders by their branches belonging to different quadrants on the DD plane and to number of intersections and return point. This makes possible studying the fine structure of any length sequences with exactly established functions (genes, intergenic space, repeat regions etc...) and comparing various locuses, as well as searching for homological regions, which can allow us to work out mathematically strong genomic signature formalism [51, 52].

We note that there exist many types of trianders. A triander corresponding to a gene we call a *genogram*, and a triander corresponding to intergenic space we call a *gapgram.* If some branches intersect each other we say about *intersecting triander*, if a branch intersects itself producing knots, we say about *knot triander.* Branches can also have (multiple) return points, and then we say about *returned triander.* Thus, the determinative degree walk "topologizing" in our sense means that we identify trianders having definite structure topological features (knot, intersection, return point) and place them into a special class.

So we may hope that such topological classification of trianders can actually help in solving by visual way the inverse problem: for a given sequence to predict

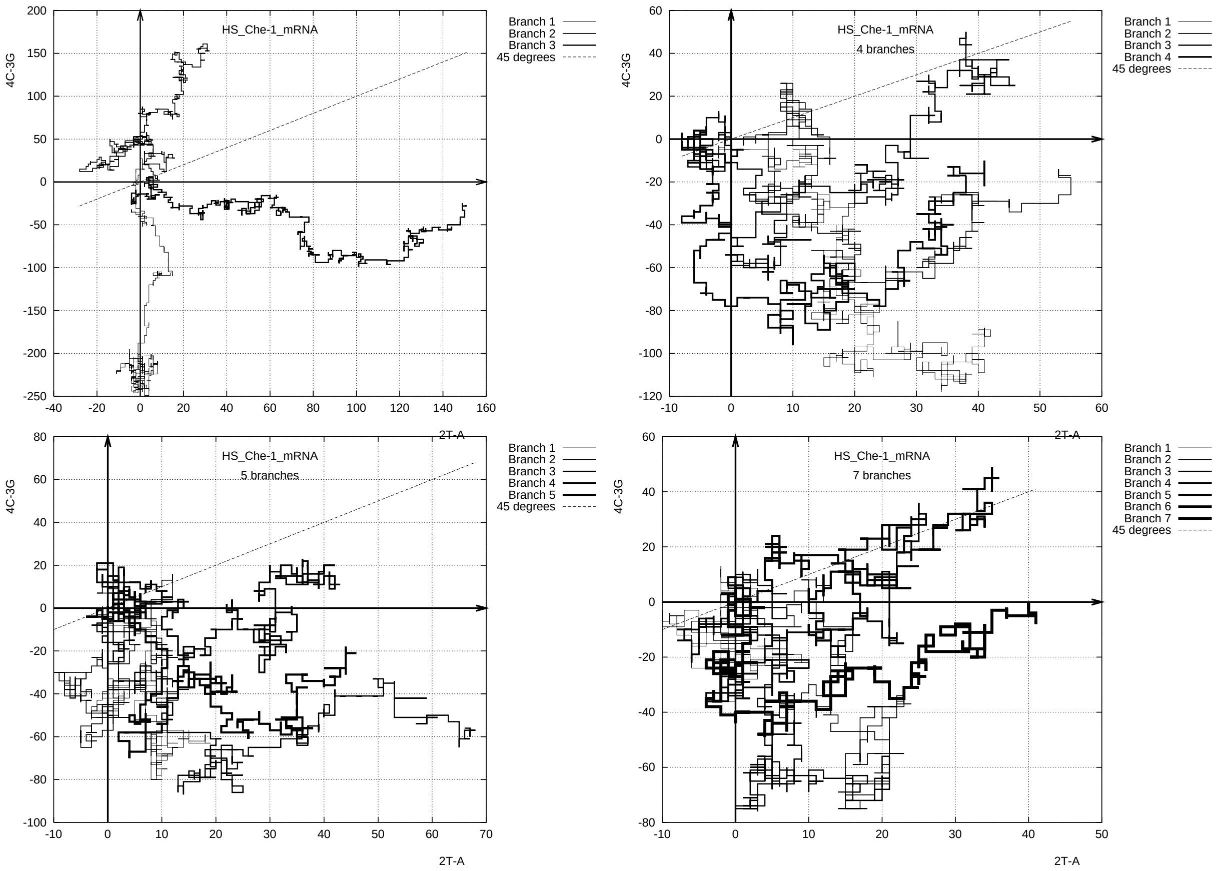

Figure 7: Trianders of the Homo sapiens Che-1 mRNA gene (first) and analogs of trianders with phases = 4,5,7 . The strong triplet signal is clearly seen.

its possible function.

Let $n_{\mathbf{X}}(i)$ be cumulative quantity of nucleotide $\mathbf{X}$ after i steps, then the DD plane quadrants are defined by (10)–(11), and therefore

I: $\mathbf{2}n_{\mathbf{T}}(i) - n_{\mathbf{A}}(i) > 0; \mathbf{4}n_{\mathbf{C}}(i) - \mathbf{3}n_{\mathbf{G}}(i) > 0;$

II: $\mathbf{2}n_{\mathbf{T}}(i) - n_{\mathbf{A}}(i) < 0; \mathbf{4}n_{\mathbf{C}}(i) - \mathbf{3}n_{\mathbf{G}}(i) > 0;$

III: $\mathbf{2}n_{\mathbf{T}}(i) - n_{\mathbf{A}}(i) < 0; \mathbf{4}n_{\mathbf{C}}(i) - \mathbf{3}n_{\mathbf{G}}(i) < 0;$

IV: $\mathbf{2}n_{\mathbf{T}}(i) - n_{\mathbf{A}}(i) < 0; \mathbf{4}n_{\mathbf{C}}(i) - \mathbf{3}n_{\mathbf{G}}(i) < 0.$

After examination of around 2000 eukaryotic and prokaryotic sequences we found all trianders can be distinguished into several types. The first type is a *chaotic triander* which has no definite branch structure, other ones can be called *ordered trianders.* To work out the general classification of ordered trianders and description of branches we introduce the notion:

$$\text{Type } \mathbf{A\text{-}B\text{-}C}_{\mathbf{F}}^{(x,y)}\,(\mathbf{E}), \qquad (19)$$

where **A** is quadrant where 1st branch lies, **B** is quadrant where 2nd branch lies,

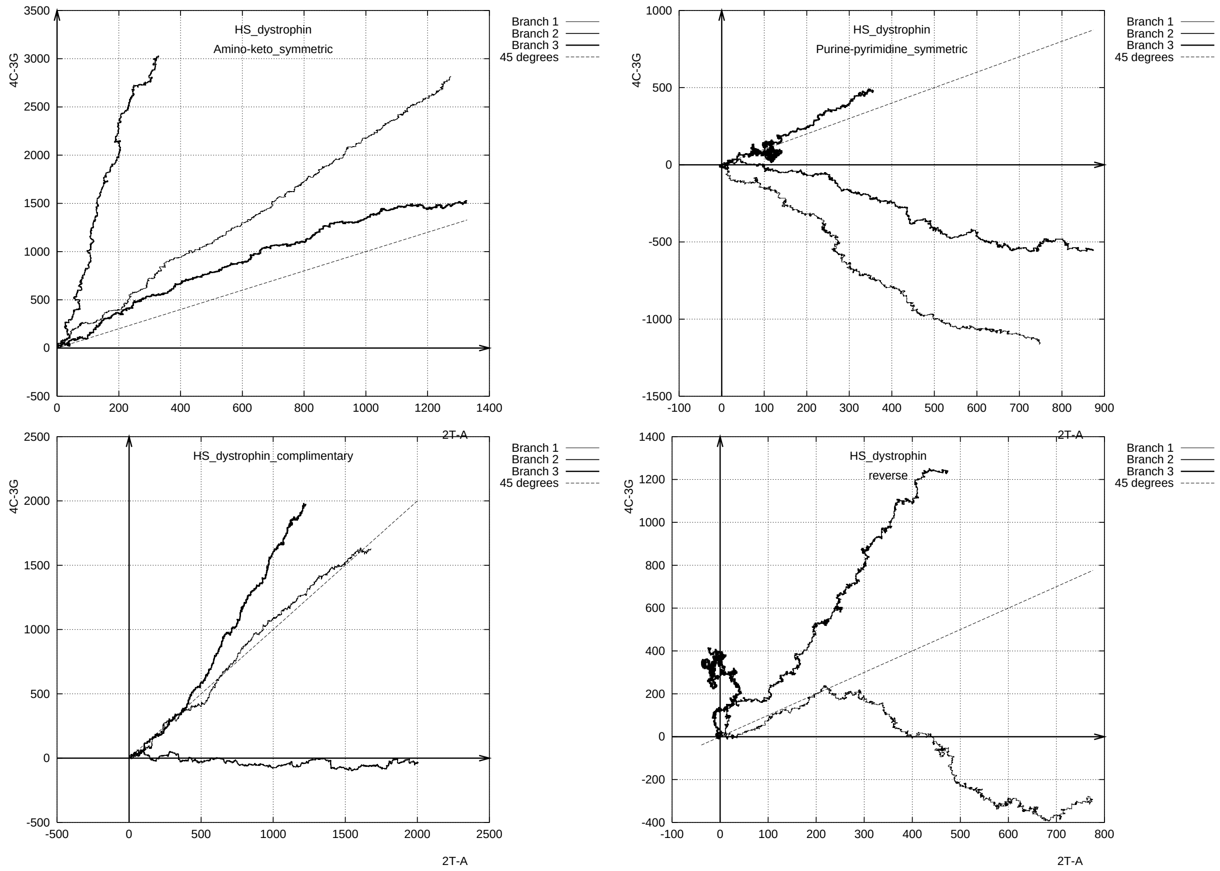

Figure 8: The trianders for the transformed sequences of Homo sapiens dystrophin (DMD), transcript variant D140ab, mRNA in case of the amino-keto symmetry, purine-pyrimidine symmetry, complementarity symmetry and reverse sequence reading. The original nontransformed triander is shown on Fig. 3.

C is quadrant where 3rd branch lies; **E** is characteristics of triander as a whole; indices **F** and (x, y) describe properties of corresponding separate branches (also for **A** and **B**) which will be explained below.

In general there are $4^3 = 64$ possible ordered triander types classified by quadrants only. We will identify trianders which differ by permutation, because it corresponds to ORF shift, thus decreasing to 24 types. Nevertheless, observation showed that there exist only 7 triander types: **I-I-I**, **I-I-II**, **I-I-III**, **I-I-IV**, **I-II-III**, **I-II-IV**, **I-III-IV**. For example, the Type **I-I-II** includes the Types **I-II-I** and Types **II-I-I**, if we shift ORF to 1 and 2, but on figures we show exact triander names (19).

If e.g. a branch crosses from **I** quadrant to **II** quadrant, we denote that by fraction **I/II**. For instance, the triander of Homo Sapiens dystrophin gene Fig. 3, is of Type **II-I-I/IV**.

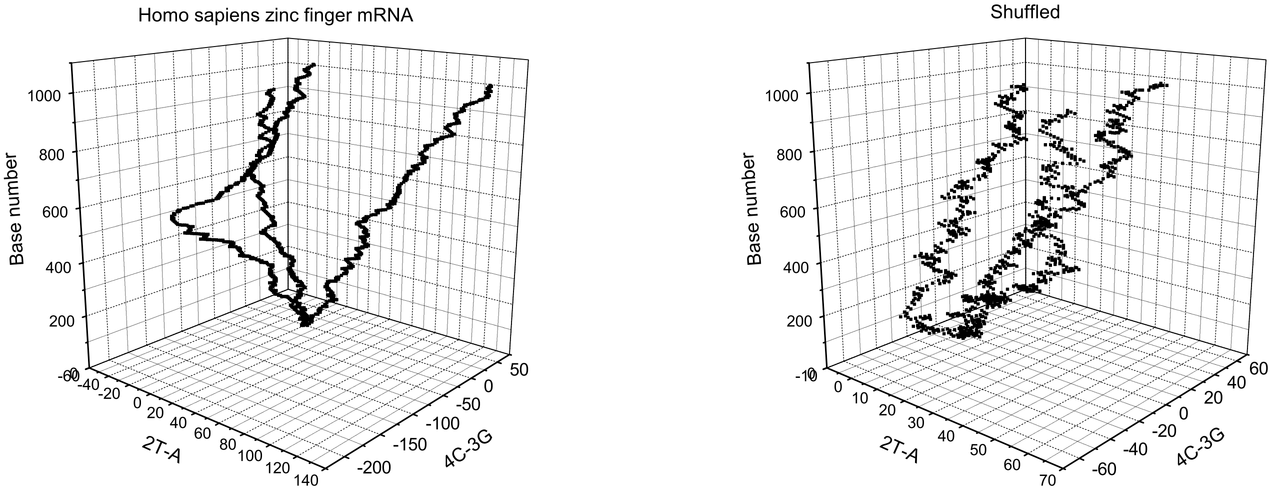

Figure 9: Three-dimensional trianders of the Homo sapiens zinc finger protein 265 (ZNF265), mRNA (left) and its shuffled version (right).

The additional qualitative features of triander as a whole observed from sequence examination are

E=sharp, flat, parallel.

For branch properties we have

x,y denotes axis to which a branch is parallel,

F=blury, loop, smooth, oscillative (horizontal, vertical).

Separately we can describe "interaction" of branches as:

1) Single intersection of **A** and **B** is denoted by sign **A#B**, which gives *intersecting triander*;

2) Multiple intersection **A** and **B** is called braiding and denoted **A** $\between$ **B**, which gives *braiding triander*. See Fig. 10.

We have thoroughly analyzed 150 sequences different by function and evolution level, and for each sequence there were also constructed 100 shuffled sequences having the same nucleotide composition, but not coinciding with the examined one. They are presented in the Tables 1,2.

For every class we show a typical triander of Fig. 11, where the following real sequences are presented:

a) Chaotic triander. Dengue virus type 1 strain FGA/NA d1d, intergenic space: AF226686.

Table 1: Examined sequences

Type	Sequence description	Number
I,I,I	Arabidopsis thaliana clone 7867 mRNA,	
	C.elegans essential lethal-805, myotactin complex	2
	C.elegans heterochronic gene LIN-42	1
	C.elegans ribosomal protein L32	1
	Dros melanogaster chromosome 3R	1
	HS apoptosis-associated tyrosine kinase	2
	HS coagulation factor VIII	1
	HS COL4A6 gene for a6(IV) collagen	2
	HS cytochrome P450,family 1-8, subfamily A,B,C,F,J	15
	HS DNA cross-link repair 1A,1C (PSO2 homolog, S. cerevisiae)	2
	HS dystrophin (DMD) gene, exons,transcript variants	4
	HS exportin, tRNA	1
	HS FGG gene for fibrinogen,	2
	HS fibrinogen alpha chain gene, complete mRNA	4
	HS glutathione synthetase (GSS), mRNA	2
	HS H19 gene,complete sequence,mRNA	2
	HS myeloid ecotropic viral integration site mRNA	1
	HS MESTIT1 antisense RNA,intergenic space	1
	HS mRNA 10q24.3-qter	1
	Hs mucosal vascular addressin cell adhesion molecule	1
	HS phenylalanine hydroxylase (PAH)	3
	Hs phytanoyl-CoA hydroxylase interacting protein	1
	HS retinal dystrophin (DMD) gene Exon 30	1
	Hs solute carrier family 22 (organic anion transporter),	1
	HS suppressor of cytokine signaling,intergenic space	1
	HS syntrophin, alpha 1	3
	HS vitelliform macular dystrophy	2
	HS, chorionic somatotropin hormone	1
	HS,genomic DNA chrom1,intregenic space	1
	HS,genomic DNA P450 intergenic space	1
	Human CYP2D7BP pseudogene for cytochrome P450 2D6	1
	Human mRNA encoding placental lactogen hormone	1
	Human nested gene protein gene	1
	Mus musculus insulin-like growth factor	2
	Mus musculus interleukin 11 receptor	1
	Mus musculus like-glycosyltransferase mRNA	1
	Mus musculus similar to mitochondrial ribosomal protein S36	2
	Rat gene for alpha-fibrinogen	1
	Rattus norvegicus cytochrome P450 IIA3 mRNA, 3' end	2
	Similar to FBJ murine osteosarcoma viral oncogene	1
	Takifugu rubripes DMD gene	1

Table 2: Examined sequences

Type	Sequence description	Number
	Dengue virus type 1 strain FGA/NA,complete genome.	1
	HS ATP-binding cassette, sub-family C (CFTR/MRP),mRNA	3
	HS aldehyde dehydrogenase 1 family	5
	Human fibrinogen beta-chain mRNA, partial cds 41b1,short asym	2
	Homo sapiens fibrinogen-like 1 (FGL1), transcript variant	1
	C.elegans immunoglobulin	1
	Hs spondyloepiphyseal dysplasia late RNA	1
	C.elegans nematode cuticle collagen	1
	C.elegans cuticle collagen family member (28.9 kD)	1
I,I,IV	Chaetosphaeridium globosum chloroplast, complete genomes	1
	C. elegans Collagen with Endostatin domain CLE-1	1
	Hs aldehyde dehydrogenase 1 family, member A2 (ALDH1A2)	1
	Hs chromosome 20 open reading frame 1 (C20orf1), mRNA	1
	HS ATP-binding cassette, sub-family C (CFTR/MRP)mRNA	2
	Human cytochrome P450 (CYP2A13) gene, complete cds.	1
	Hs cytochrome P450 2S1 (CYP2S1) mRNA, complete cds.	1
	Hs P450 (cytochrome) oxidoreductase (POR),	1
	Hs cytochrome P450, family 2,39,51,20	4
	Hs chromosome 1 MRG1 intergenic space	1
	Hs cytochrome P450 intergenic space	1
I,II,IV	HS dystrophin (DMD) D140ab, variants,mRNA.	5
	H.sapiens mRNA for ribosomal protein L30	1
	Human mRNA for ribosomal protein L32	1
I,I,III	Hs collagen, type IX, alpha 2 (COL9A2), mRNA	1
	Mus musculus ribosomal protein L32	1
	Mus musculus, ribosomal protein L30 BC002060	1
	Homo sapiens apoptosis antagonizing transcription factor (AATF)	5
I,II,III	Homo sapiens H1F5 histone family	1
	Hs cytochrome P450, family 17, subfamily A, polypeptide 1	1
	HS genomic cluster,H1 histone family, member 5	1
	HS survival of motor neuron 1, telomeric (SMN1)	2
	HS Che-1 mRNA, complete cds.123mRNA	1
	Hs ZIS1protein 265 (ZNF265) mRNA,	1
I,III,IV	Hs utrophin (homologous to dystrophin)3bls11	1
	Macaca fascicularis RPL30 mRNA, family	4
I,I,II	HS bestrophin (VMD2) mRNA, alternatively spliced product	1
	HS BTG family, member 2 (BTG2), mRNA	1
	HS Cbp/p300-interacting transactivator,with Glu/Asp-rich	1
	Human msg1-related gene 1 (mrg1) mRNA	1

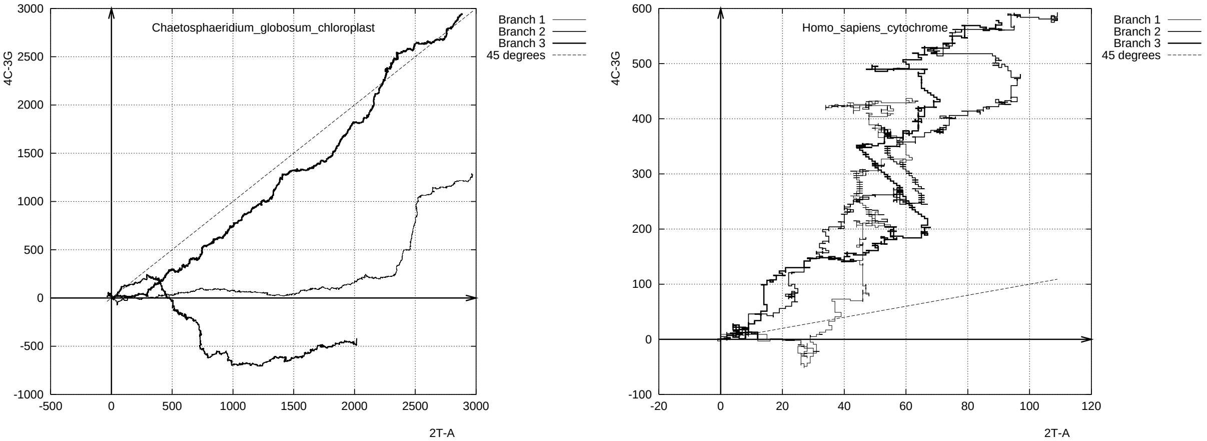

Figure 10: Intersecting triander, Chaetosphaeridium globosum chloroplast, complete genome (left) and braiding triander Homo sapiens cytochrome P450 2f1 (CYP2F1P) (right).

b) Type **I-I-I**^{y}. Homo sapiens cytochrome P450, family 2, subfamily F, polypeptide 1 (CYP2F1), mRNA: NM_000774.

c) Type **II**$_{blury}$**-I**^{x}**-I**^{y}(flat). Homo sapiens Cbp/p300-interacting transactivator, mRNA: NM_006079.

d) Type **III**$_{oscill}$**-I**$_{blury}$**-I**$_{oscill}$. Homo sapiens collagen, type IX, alpha 2 (COL9A2), mRNA: NM_001852.

e) Type **IV-I#I**. Caenorhabditis elegans immunoglobulin domain-containing protein family member (106.4 kD), mRNA: NM_171617.

f) Type **III-II-I**^{x}(sharp). Homo sapiens H1 histone family, member 5 (H1F5), mRNA: NM_005322.

g) Type **II#I-IV**. Homo sapiens dystrophin (muscular dystrophy, Duchenne and Becker types) (DMD), transcript variant D140ab, mRNA: NM_004022.

h) Type **I**$_{loop}$**-I-IV**. Homo sapiens utrophin (homologous to dystrophin) (UTRN), mRNA: NM_007124.

Further more careful topological classification and analysis of two- and three-dimensional trianders can be made using some of the topological curves methods [53, 54, 55] or the knot theory (see e.g. [56, 57]).

6 Conclusions

We can conclude that the introduced determinative degree DNA walk method confirms the "mosaic" stricture of genome, shows parts with different nucleotide content and "strength", and so allows us to find the "fine structure" of nucleotide sequences.

We propose a general method of identification of DNA sequence “by triander”, which can be treated as a unique “genogram”, “gene passport”, etc. The two- and three-dimensional trianders are introduced and their features are studied.

The difference of the nucleotide sequences fine structure in genes and the intergenic space is shown. Also there is a clear triplet signal in coding locuses which is absent in the intergenic space and is independent from the sequence length, but depends from composition only. All plots are compared with corresponding shuffled sequences of the same nucleotide composition, which allows us to extract real ordering effect from composition influence.

We have constructed the classification of trianders, on its basis a detail working out signatures of functionally different genomic regions can be made.

Acknowledgments. We would like to thank S. Cebrat and M. R. Dudek for kind hospitality at SmORFland (Inst. Microbiology, Wroclaw) and sharing their useful DNA walk experience. Also we are grateful to A. Yu. Berezhnoy, N. A. Chashchin, A. A. Gusakov, V. Knecht, G. Ch. Kourinnoy, P. Mackiewicz, B. V. Novikov, L. A. Livshits, S. Sachse, Yu. G. Shckorbatov, O. A. Tretyakov for very fruitful discussions, and G. Findley, H. Grubmüller, P. Jarvis, D. C. Torney, C. Zhang for sending their interesting papers. We are thankful to S. M. Donets for language checking and V. Kalashnikov for creation of original special programs for sequence analysis and indispensable assistance in the related GNUPLOT/PERL/C programming.

References

[1] D. C. Torney, C. C. Whittaker, and G. Xie, *The statistical properties of human coding sequences*, J. Mol. Biol. **286** (1999), 1461–1469.

[2] M. Bulmer, *A statistical analysis of nucleotide sequences of introns and exons in human genes*, Mol. Biol. Evol. **4** (1987), 395–405.

[3] L. Luo, W. Lee, L. Jia, F. Ji, and L. Tsai, *Statistical correlation of nucleotides in a DNA sequence*, Phys. Rev. **E58** (1998), 861–871.

[4] K. Nieselt-Struwe, *Graphs in sequence spaces: A review of statistical geometry*, Biophys. Chem. **66** (1997), 111–131.

[5] J. W. Fickett, D. C. Torney, and D. R. Wolf, *Base compositional structure of genomes*, Genomics **13** (1992), 1056–1064.

[6] S. V. Buldyrev, N. V. Dokholyan, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, and G. M. Viswanathan, *Analysis of DNA sequences using methods of statistical physics*, Physica **A249** (1998), 430–438.

[7] M. Y. Azbel, *Universality of DNA statistical structure*, Phys. Rev. Lett. **75** (1995), 168–171.

[8] G. L. Findley, A. M. Findley, and S. P. McGlynn, *Symmetry characteristics of the genetic code*, Proc. Natl. Acad. Sci. USA **79** (1982), 7061–7065.

[9] J. E. M. Hornos and Y. M. M. Hornos, *Model for the evolution of the genetic code*, Phys. Rev. Lett. **71** (1993), 4401–4404.

[10] J. D. Bashford, I. Tsohantjis, and P. D. Jarvis, *Codon and nucleotide assignments in a supersymmetric model of the genetic code*, Phys. Lett. **A233** (1997), 481–488.

[11] T. Bhry, A. Cziryk, T. Vicsek, and B. Major, *Application of vector space techniques to DNA*, Fractals **6** (1998), 205–210.

[12] M. Forger and S. Sachse, *Lie superalgebras and the multiplet structure of the genetic code* **I***: Codon representations*, *Inst.* de *Mat.* e *Estat. preprint*, `math-ph/9808001`, Sao Paulo, 1998, 23 p.

[13] L. Frappat, A. Sciarrino, and P. Sorba, *A crystal base for the genetic code*, Phys.Lett. **A250** (1998), 214–221.

[14] Y. Nakamura, T. Gojobori, and T. Ikemura, *Codon usage tabulated from international DNA sequence databases: Status for the year 2000*, Nucl. Acds Res. **28** (2000), 292.

[15] E. Hamori, *Novel DNA sequence representations*, Nature **314** (1985), 585–586.

[16] M. A. Gates, *Simpler DNA sequence representations*, Nature **316** (1985), 219.

[17] C. L. Berthelsen, J. A. Glazier, and M. H. Skolnick, *Global fractal dimension of human DNA sequences treated as pseudorandom walks*, Phys. Rev. **A45** (1992), 8902–8913.

[18] M. Y. Azbel, *Random two-component one-dimensional ising model for heteropolymer melting*, Phys. Rev. Lett. **31** (1973), 589–592.

[19] J. R. Lobry, *A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria*, Biochimie **78** (1996), 323–326.

[20] G. Bernardi, B. Olofsson, and J. Filipski, *The mosaic genome of warm-blooded vertebtates*, Science **228** (1985), 953–958.

[21] S. Cebrat and M. R. Dudek, *The effect of DNA phase structure on DNA walks*, Eur. Phys. J. **3** (1998), 271–276.

[22] M. Kowalczuk, P. Mackiewicz, and D. Mackiewicz, *DNA asymmetry and replicational mutational pressure*, J. Appl. Genet. **42** (2001), 553–577.

[23] D. Duplij and S. Duplij, *Symmetry analysis of genetic code and determinative degree*, Biophysical Bull. Kharkov Univ. **488** (2000), 60–70.

[24] D. Duplij, S. Duplij, and N. Chashchin, *Symmetric properties of genetic code*, Biopolymers and Cell **16** (2000), 449–454.

[25] M. Yčac, *The biological code*, North-Holland, Amsterdam, 1969.

[26] B. Lewin, *Genes*, Wiley and Sons, New York, 1983.

[27] G. Stent and R. Kalindar, *Molecular genetics*, Mir, M., 1981.

[28] M. Singer and P. Berg, *Genes and genomes*, University Science Books, Mill Valley, 1991.

[29] U. B. Rumer, *Sistematics of codons in the genetic code*, DAN SSSR **183** (1968), 225–226.

[30] V. A. Karasev and S. G. Sorokin, *Topological structure of the genetic code*, Genetika **33** (1997), 744–751.

[31] V. A. Karasev, *Rhombic version of genetic vocabulary based on complementary of encoding nucleotides*, Vest. Leningr. un-ta **1** (1976), 93–97.

[32] V. A. Ratner, *Structure and evolution of the genetic code*, in Itogi nauki i tekhniki. Ser. Mol. Biol., Vol. 21, VINITI, M., 1985, pp. 158–197.

[33] ———, *Genetic code as a system*, Soros Educational J. **6** (2000), 15–22.

[34] U. D. Rumer, *On codon sistematics in the genetic code*, DAN SSSR **187** (1969), 937–938.

[35] D. Duplij and S. Duplij, *Determinative degree and nucleotide content of DNA strands*, Biophysical Bull. Kharkov Univ. **525** (2001), 86–92.

[36] V. V. Sukhodolec, *A sence of the genetic code: reconstruction of the prebiologocal evolutin stage*, Genetika **21** (1985), 1589–1599.

[37] S. Y. Maslov, *On the nature of biological code and its possible evolution*, Biophysics (Moscow) **26** (1981), 632–635.

[38] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M. R. Dudek, and S. Cebrat, *High correlation between the turnover of nucleotides under mutational pressure and the DNA composition*, BMC evolutionary biology **17** (2001), 1–13.

[39] M. Dudek, S. Cebrat, M. Kowalczuk, P. Mackiewicz, A. Nowicka, D. Mackiewicz, and M. Dudkiewicz, *Information weights of nucleotides in DNA sequences*, *Inst. Microbiology preprint*, cond-mat/0301371, `cond-mat/0301371`, Wroclaw, 2002, 8 p.

[40] N. V. Zheltovsky, S. A. Samoilenko, and D. N. Govorun, in Spectroscopy of Biological Molecules, Societa Editrice Esculapio, Bologna, 1989, pp. 159–172.

[41] D. N. Govorun, V. D. Danchuk, Y. R. Mishchuk, I. V. Kondratyuk, N. F. Radomsky, and N. V. Zheltovsky, *AM1 calculation of the nucleic acid bases structure and vibrational spectra*, J. Mol. Structure **267** (1992), 99–103.

[42] G. M. Ziegler, *Lectures on Polytopes*, Springer-Verlag, Berlin, 1995.

[43] C. T. Zhang, *A symmetrical theory of DNA sequences and its applications.*, J. Theor. Biol. **187** (1997), 297–306.

[44] J. Sponer, J. Leszczynski, V. Vetterl, and P. Hobza, *Base stacking and hydrogen bonding in protonated cytosine dimer: The role of molecular ion-dipole and induction interactions*, J. Biomolecular Structure and Dynamics **13** (1996), 695–705.

[45] B. Schneider and H. B. Berman, *Hydration of DNA bases is local*, Biophysical J. **69** (1995), 2661–2669.

[46] C. Wu, *DNA strand asymmetry*, Nature **352** (1991), 114.

[47] M. P. Francino and H. Ochman, *Strand asymmetries in DNA evolution*, Trends Genet. **13** (1997), 240–245.

[48] N. Sueoka, *Intrastrand parity rules of dna base composition and usage biases in synonymous codons*, J. Mol. Evol. **40** (1995), 318–325.

[49] M. Yagi, Y. Takeshima, H. Wada, H. Nakamura, and M. Matsuo, *Two alternative exons can result from activation of the cryptic splice acceptor site deep within intron 2 of the dystrophin gene in a patient with as yet asymptomatic dystrophinopathy*, Hum. Genet. **267** (2003), 164–170.

[50] S. Skiena, *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*, Addison-Wesley, Reading, 1990.

[51] Z.-B. Wu, *Self-similarity limits of genomic signatures*, *Inst. Mechanics preprint*, `cond-mat/0212091`, Beijing, 2002, 12 p.

[52] S. Bergmann, J. Ihmels, and N. Barkai, *Self-similarity limits of genomic signatures*, *Weizmann Inst. Science preprint*, `cond-mat/0210038`, Rehovot, 2002, 12 p.

[53] I. G. Petrovskiy, *On the topology of real plane algebraic curves*, Ann. Math. **39** (1938), 189–209.

[54] V. A. Rokhlin, *Complex orientation of real algebraic curves*, Func. Anal. Appl. **8** (1974), 71–75.

[55] V. I. Arnold and O. A. Oleinik, *Topology of real algebraic manifolds*, Vestnik Mosk. Univ., Ser. I, Mat. i Mekh **A249** (1979), 7–17.

[56] V. G. Turaev, *Quantum Invariants of Knots and 3-Manifolds*, W. de Greuter, Berlin, 1994.

[57] L. H. Kauffman, *Knots and Physics*, World Sci., Singapure, 1991.

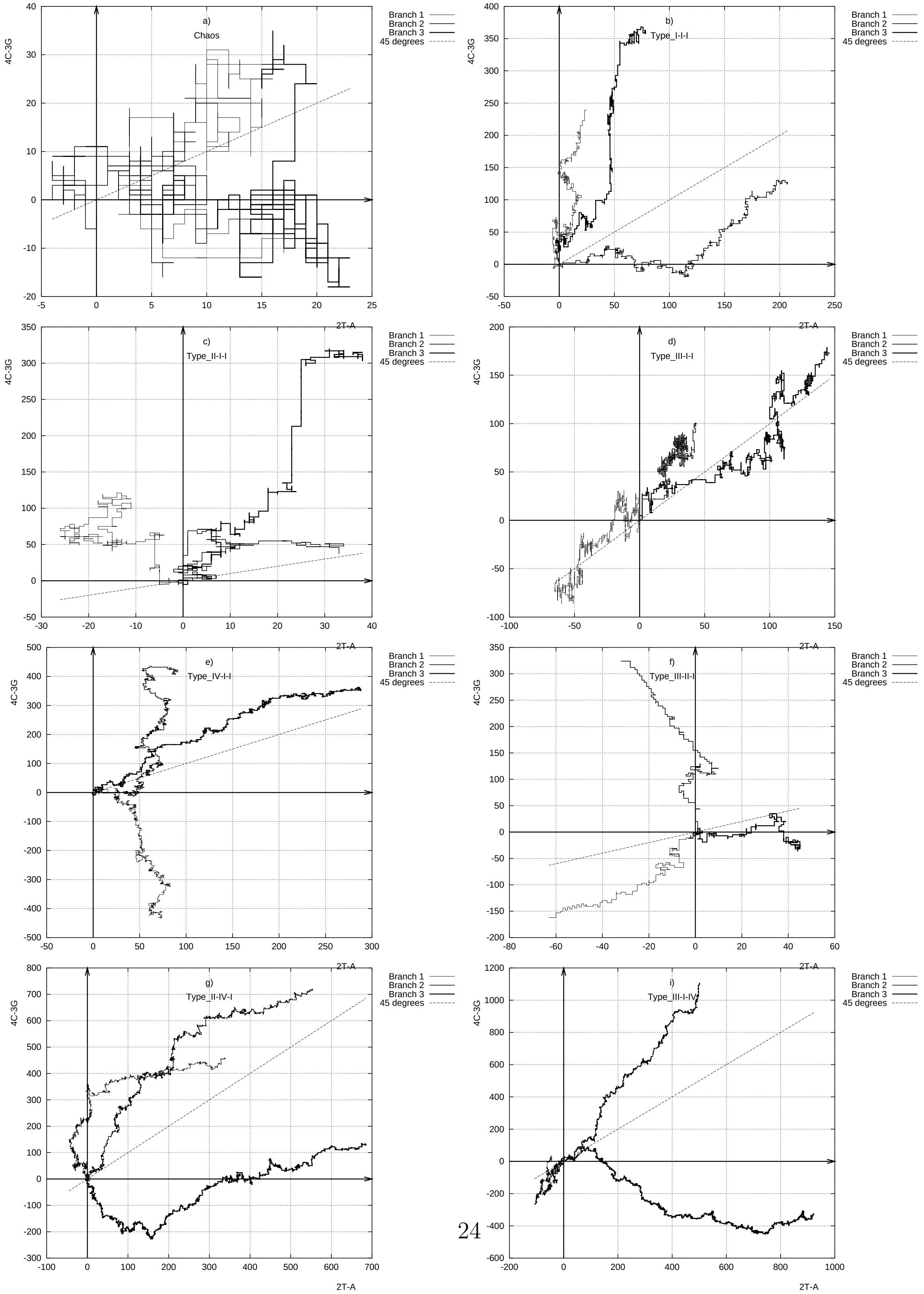

Figure 11: Classification of trianders.